A DATA SEGMENTATION APPROACH TO MODELING FRESHMAN RETENTION

Sarah Caro, PhD
University of New Haven

Introduction

- About me:
  - *Sarah Caro, PhD, Cognitive Psychology*
  - *Senior Research Analyst, Institutional Research, University of New Haven (2014-present)*

- About UNH
  - *Connecticut Private, Non-Profit,*
    - 4500 UG and 1500 Graduate
    - Freshmen classes 1200 – 1400
      - *Fall-to-Fall retention ~ 79%*

Modeling Dataset

- Fall-to-Fall Freshman Retention
  - *3-year training dataset: F14, F15, and F16*

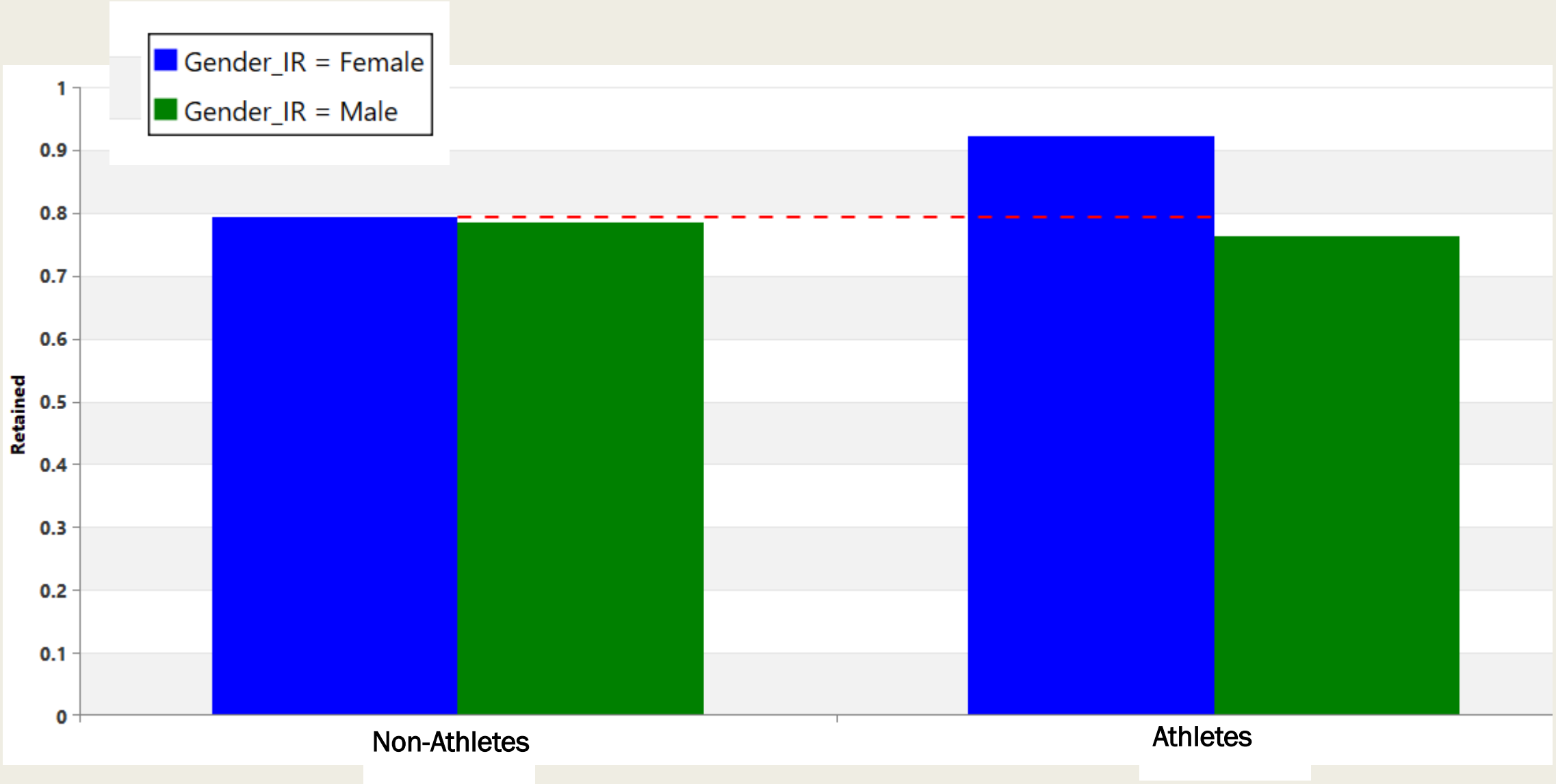
- Early term and EOT (End-of-Term) models
  - *Early Term: Reach out to at-risk students*
  - *EOT: Get a predicted percent of freshmen returning sophomore fall* ✓

- Variables
  - *Academic*
    - SAT Scores, HS GPA;
    - EOT GPA, EOT Credits Earned, # High DFW courses
  - *Demographic/Financial*
    - Gender, Ethnicity, Home State, Distance from Campus; Pell Status, Family income, EFC
  - *Engagement/Interest*
    - Admissions Events, LLC membership, Rec Center Visits, Course Eval Completion Rate, Course Eval responses; Date of Application; Tuition Deposit Month

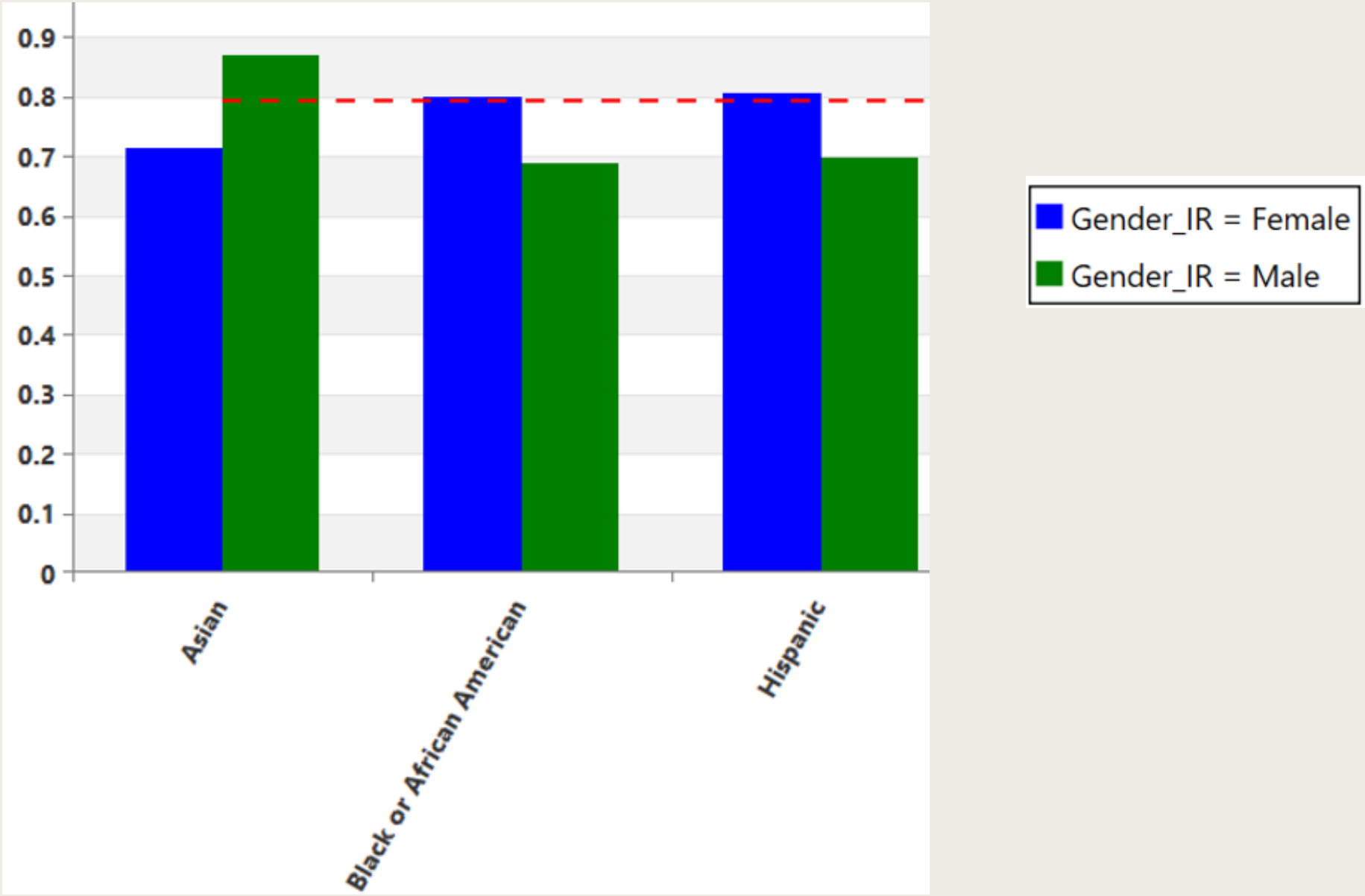
Modeling Approach

- Traditional approach: Clean and vet dataset, then create one model based on entire dataset
- New approach: Separate models for distinct segments of freshman population
- Different retention trends by gender?

Some Gender Trends: Athlete Status



Some Gender Trends: Ethnicity



Some Gender Trends: Engagement/Interest Variables

Predictor	Female student Correlation with Retention	Male student Correlation with Retention
# Admissions Events	0.12	0.22
# Rec Center Visits	0.06	0.08
LLC Membership	0.09	0.16
Days Btwn App Date & Semester Start	0.11	0.15

Segment the Dataset?

- Leave more “room” in model for targeted variables

- How to split the dataset?
  - *Male vs. Female*
  - *Residents vs. Commuters*

- Profiling to find differences between segments

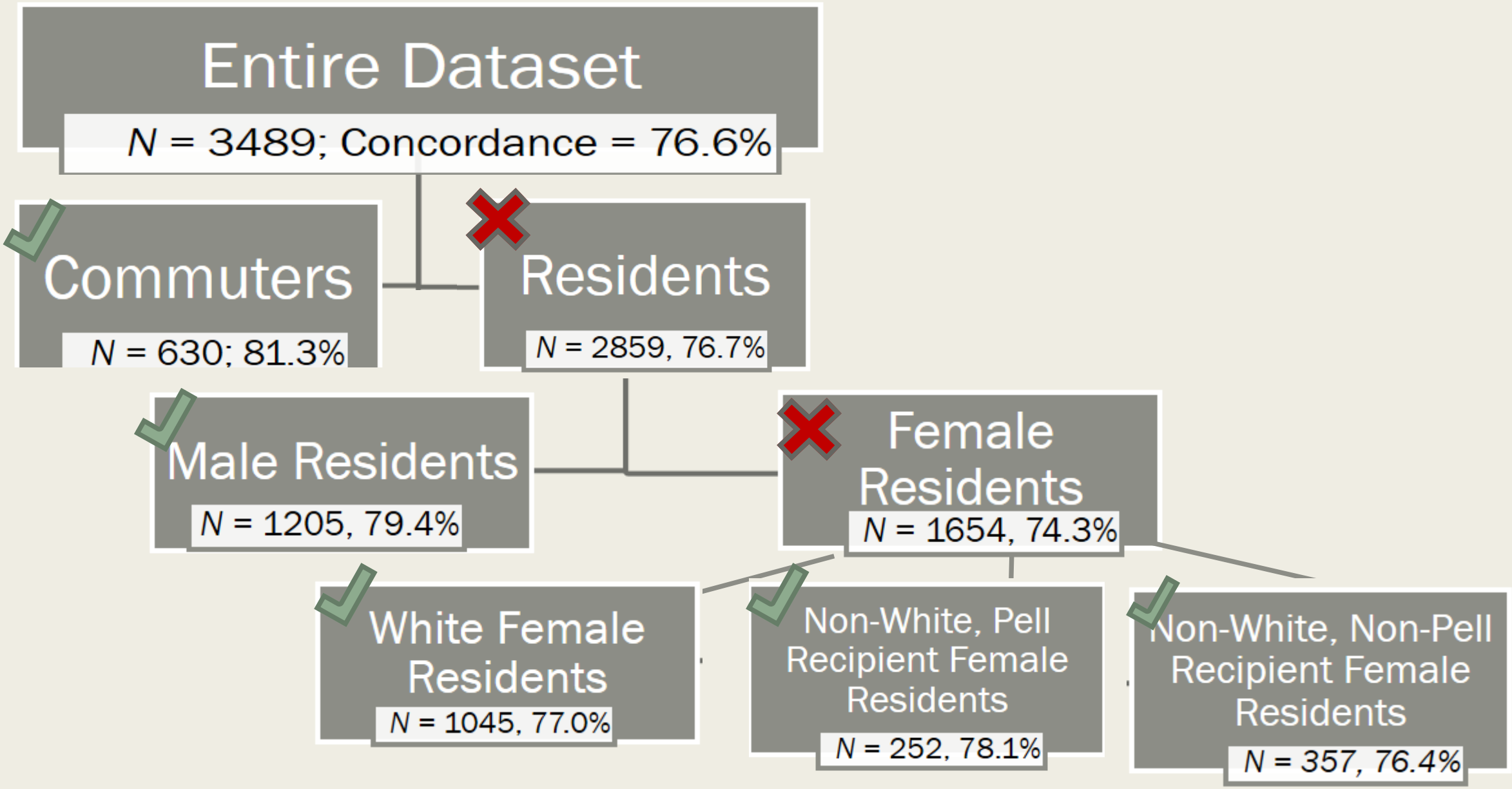
Comparing Segments: Residents/Commuters

- Comparatively, **RESIDENTS** (Profiling, *p < .01)*:
  - *Financial: Higher EFC and family income*
  - *Engagement/Interest: More Admissions events, Apply and Deposit Earlier, course eval completion*
  - *Geographical:* Come from farther; fewer CT residents
  - *Academic Quality: Higher on almost all measures of academic quality.*
  - *Areas of Study: More A&S and HCLC; fewer Engineering*

Comparing Segments: Male/Female Students

- Comparatively, FEMALE STUDENTS (Profiling, $p < .05$):
  - *Academic Quality:*
    - Pre-admission: Higher HS GPA, Higher SAT Verbal
      - *Lower SAT Math*
    - At UNH: More High-DFW Credits, More Credits taken/earned, Higher UNH Term GPA
  - *Engagement/Interest*
    - Apply and Deposit Earlier, More Course Evals; Living on Campus
      - *(Fewer Rec Ctr visits)*
  - *Geographical: Fewer CT Residents; more MA residents*
  - *Areas of Study: More Forensic Science, CJ and Psych; Fewer Engineering and Business*

Segmenting the Dataset: Find best concordance



Notable model differences

- Common predictors: EOT GPA, # Credits Earned
- Commuters
  - *2nd Highest Contributor: "West Haven High School Scholarship"*
- Male Residents:
  - *3rd Highest Contributor: LLC status*
- White Female Residents
  - *Athlete flag – 4th highest*
  - *No financial vars*
- Non-White Female **Pell** Recipients
  - *No financial vars*
  - *Measures of engagement (early applications and completing course evals)*
- Non-White Female Non-Pell Recipients
  - *Financial Vars like EFC and Need*

Testing the Models

- How well do these models predict retention on a new dataset?
  - *Mid-May Registration: (pretty good indicator of sophomore fall retention)*

- **Prior years:** Multiply the number registered mid-May by .953

- **F17 (test) Cohort:** 1125 students are registered Mid-May
  - *1125 x .953 = 1072*
  - *76.1% of the original freshman cohort.*

Scoring the models

- Score models – get a predicted percent returning for the *test* dataset.

Model	Group	Test N (F17 Freshmen)	Predicted Percent Returning	Predicted Number Returning
1-Model	All Students	1408	75.4%	1062

Scoring the models

- Score models – get a predicted percent returning for the test dataset.

Model	Group	Test N (F17 Freshmen)	Predicted Percent Returning	Predicted Number Returning
5-Model	Commuters	297	69%	206
	Male Residents	488	77%	374
	Fem White Residents	358	81%	291
	Fem Non-White Non-Pell Residents	143	80%	114
	Fem Non-White Pell Residents	122	70%	85
	Combination of 5 Models	1408	75.9%	1069

Data Segmentation – Worth the Trouble?

- Predicted Retention Rates:
  - *Mid-May Trend Estimate: 76.1%*
  - *5-Model Method: 75.9%*
  - *1-Model method: 75.4%*

- 1-Model method concordance: 76.6%
  - *5-Model method concordance: 78.6%*

- Theoretically, you can hone in on variables tailored to more “homogenous” groups.

Questions/Discussion?

- Questions/Discussion

- Contact:
  - *Sarah Caro, scaro@newhaven.edu*